Eur Spine J (2012) 21:1590–1595
DOI 10.1007/s00586-012-2169-3

ORIGINAL ARTICLE

Analysis of internal construct validity of the SRS-24 questionnaire

Dominique A. Rothenfluh · Georg Neubauer · Juergen Klasen · Kan Min

Received: 27 April 2011 / Revised: 21 November 2011 / Accepted: 19 January 2012 / Published online: 8 February 2012

Abstract
Purpose The SRS-24 questionnaire was originally validated using methods of classical test theory, but internal construct validity has never been shown. Internal construct validity, i.e. unidimensionality and linearity, is a fundamental arithmetic requirement and needs to be shown for a scale for summating any set of Likert-type items. Here, internal construct validity of the SRS-24 questionnaire in adolescent idiopathic scoliosis (AIS) patients is analyzed.
Methods 232 SRS-24 questionnaires distributed to 116 patients with AIS pre-operatively and at postoperative follow-up were analyzed. 103 patients were females; the average age was 16.5 ± 7.1 years. The questionnaires were subjected to Rasch analysis using the RUMM2020 software package.
Results All seven domains of the SRS-24 showed misfit to the Rasch model, and three of seven were unidimensional. Unidimensionality and linearity could only be achieved for an aggregate score by separating pre- and postoperative items and omitting items which caused model misfit. Reducing the questionnaire to six pre-operative items ($p = 0.098$; 2.25% t tests) and five postoperative items ($p = 0.267$; 3.70% t tests) yields model fit and unidimensionality for both summated scores. The person-separation indices (PSI) were 0.67 and 0.69, respectively, for the pre- and postoperative patients.
Conclusions The SRS-24 score is a non-linear and multidimensional construct. Adding the items into a single value is therefore not supported and invalid in principle. Making profound changes to the questionnaire yields a score which fulfills the properties of internal construct validity and supports its use a change score for outcome measurement.

Keywords SRS-24 · Rasch analysis · Internal construct validity

Electronic supplementary material The online version of this article (doi:10.1007/s00586-012-2169-3) contains supplementary material, which is available to authorized users.

D. A. Rothenfluh (✉) · G. Neubauer · J. Klasen · K. Min
Department of Orthopaedics, Balgrist Clinic,
University of Zurich, Forchstrasse 340,
8008 Zurich, Switzerland
e-mail: dominique.rothenfluh@mac.com

Introduction

The Scoliosis Research Society 24-item questionnaire (SRS-24) was developed as a measure of patient satisfaction for evaluation and monitoring of patients with idiopathic scoliosis [3]. It is intended to be used as a summated score of seven dimensions comprising pre- and postoperative items and was validated using classical test-theory demonstrating reliability and external validity. Scores fulfilling the "traditional" psychometric properties have key clinically important limitations which potentially restrict their use in research as well as clinical practice. Patient-reported outcome questionnaires developed using classic test-theory yield ordinal data derived from either Likert-type or VAS scales. They are counts of numbers of responses to different questions and do not necessarily correspond to a clinically meaningful difference between the response options. A change or difference of one point may therefore vary in its meaning across the scale for every question and for every patient. It has been reported that the meaning of a 1-point change in an ordinal scale may vary up to 15-fold across the scale range and that this variation is dependent on the scale [11]. In Fig. 1, a ruler is shown

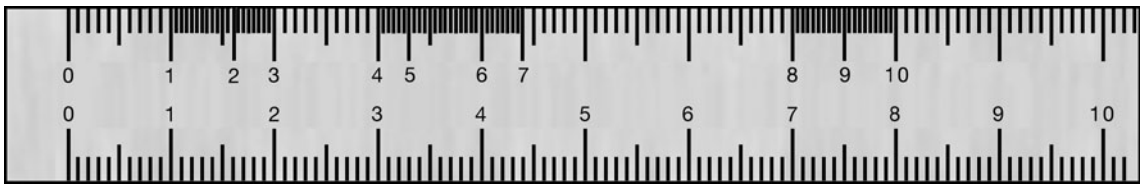

Fig. 1 Ruler: the *ruler* indicates an ordinal scale in the *top row* and an interval scale in the *bottom row*. The *top row* corresponds to ordinal observed data and the *bottom row* to interval-level latent data and demonstrates that arithmetic operations are only valid with interval-level data

which indicates an ordinal scale in the top row and an interval scale in the bottom row. Summing up numbers in the top row graphically demonstrates that adding for example 1 and 2, which may represent Likert-type response options for example, does not equal 3, whereas linear numbers meet this arithmetic requirement. In order to calculate change scores from linear data, a method is therefore needed which makes the transition from the top to the bottom row of the ruler.

Rasch measurement as well as item response theory methods is being increasingly used in patient-reported outcome measures. While external validity has been demonstrated for the SRS-24 questionnaire using methods of classical test theory [3], it has never been validated by Rasch analysis, which is currently the accepted method and "gold standard" for calibration of questionnaires and scores for outcome measurement [2, 4]. Validation of a questionnaire using Rasch analysis provides a means of making sure that a scale yields a linear score derived from ordinal scores and that it is strictly unidimensional. This allows for the legitimate calculation of a total score and measuring clinical change. When data fit the Rasch model, the questionnaire possesses internal construct validity, which comprises linearity and unidimensionality of the condition being assessed such as adolescent idiopathic scoliosis.

In the present study, the SRS-24 questionnaire is subjected to Rasch analysis to test for internal construct validity in adolescent idiopathic scoliosis (AIS) patients. In order to obtain fit to the Rasch model and therefore ensure internal construct validity, fundamental changes had to be made to the questionnaire.

Methods

Patients

A German translation of the original SRS-24 questionnaire was distributed to patients scheduled for surgery for adolescent idiopathic scoliosis (AIS) preoperatively and at 24-month postoperative follow-up. 232 questionnaires were collected for analysis from 116 consecutive patients having had surgery for AIS and of which both preoperative and follow-up questionnaires were available. Selection of the questionnaires was therefore random and depended on the availability of both for each patient included. None of the patients had any comorbidities. The average age was 16.5 ± 7.1 years. Out of the 116 patients, 103 were females (88.7%) and 13 males (11.3%). 97 of 116 patients underwent anterior correction and fusion, whereas 12 patients had posterior correction and fusion and 7 patients a combined anterior/posterior procedure. Data are summarized in Table 1.

Table 1 Patient characteristics

n total	116
Age (SD)	16.5 (7.1)
Female	103 (88.7%)
Male	13 (11.3%)
Follow-up	24 months
Anterior	97
Posterior	12
Combined	7

Rasch analysis

Analysis of the raw scores and fitting the data to the Rasch model has been described in detail before and is briefly summarized [6, 8]. The RUMM2020 software (RUMM Laboratory, Perth, Australia) was used to test Rasch model fit and unidimensionality. Fit to the model is determined by calculating item–person interaction statistics. An additional item-trait statistic tests the property of invariance across the trait as a χ^2 statistic. Misfit to the Rasch model is investigated by individual person and item fit statistics. For the individual item fit, the overall χ^2 statistic for each item is calculated, significant values indicate misfit of the individual item to the model. To take account of multiple testing, Bonferroni corrections are applied to adjust the χ^2 p value [1]. As an estimate of internal consistency, RUMM2020 calculates a person separation index (PSI) where the estimates on the logit scale for each person are used for calculation.

For a good fitting model, respondents with high levels of the attribute being measured would endorse high scoring responses, while individuals with low levels of the attribute would consistently endorse low probability curve scoring responses for each of the items. Response options for each item therefore need to be ordered. Responses to an item may reveal disordered response options as a source of item misfit.

Another factor which may affect model fit and yield wrong person estimates is an item bias known as differential item functioning (DIF). This occurs when different groups within the sample respond in a different manner to an individual item, for example. males and females, pre-operative and post-operative responses or patients doing sports versus patients not doing sports. The presence of DIF is detected by analysis of variance (ANOVA) for each item comparing scores across each level of the person factor and across different levels of the trait. DIF is indicated by a significant main effect for the person factor or by a significant interaction effect.

Internal construct validity comprises linearity, i.e. Rasch model fit, and unidimensionality. Unidimensionality is a fundamental requirement of internal construct validity [9] and needs to be shown for a scale for summating Likert-type responses into a total score [7, 11]. Unidimensionality requires that a scale is only measuring one underlying concept and is investigated by testing for multidimensionality at each level of the analysis for model fit. Testing for multidimensionality uses independent t tests to probe person estimates of potentially contrasting subsets of questions within the score. If the person estimate is found to differ between the subsets this would indicate multidimensionality of the scale. Unidimensionality is supported if the independent t test is significant (with binomial confidence intervals for a proportion) in less than 5% of the cases of the whole sample size.

For comparison of pre- and postoperative scores one-way ANOVA in the RUMM2020 software package was used.

Results

Fit of the SRS-24 questionnaire and its domains to the Rasch model

Rasch analysis of the full SRS-24 questionnaire as originally introduced revealed a non-linear construct as indicated by the model misfit ($p < 0.000001$) and a multidimensional score (11.94% of t tests significant) (Table 2). The ordinal raw scores of the SRS-24 items therefore do not fulfill the requirements of internal construct validity for summating the items into a total construct, which necessitated a more detailed analysis of the questionnaire.

In 11 out of the 24 questions disordered response options were discovered (items 1, 2, 4, 7, 8, 12, 14, 15, 20, 23, 24; numbers refer to the items in the SRS-24 questionnaire as published by Haher et al. [3]). In these items, patients were not able to distinguish between the response options offered by the respective item and their use was inconsistent with the trait being measured such that low or high levels of the attribute being measured do not necessarily endorse low or high scoring responses. Items were rescored in RUMM2020 by reducing response options to result in an item with less response options. For example, in item 2, the original response structure of 01234 with 5 Likert-Type response options which are used to calculate the score had to be rescored to 01223. This means that the response options 2 and 3 had to be collapsed to result in an item with four response options instead of five. How items were rescored is given in Table 4 for items which are kept to calculate the pre- and postoperative summated scores.

After rescoring the items with disordered response options, they were grouped into their seven domains for individual analysis of each domain of the SRS-24 questionnaire. Out of the seven domains, all were considered to show model misfit and therefore lack linearity (Table 2). Only the domain "general self-image" showed no significant deviation from the Rasch model just at the 0.05 significance level with $p = 0.0516$ and therefore borderline

Table 2 Rasch model fit statistics for the SRS-24 domains

Analysis	Item fit residual		Person fit residual		Chi square interaction		PSI	t tests (CI)	n
	Mean	SD	Mean	SD	Value	p			
Full SRS-24	0.016	1.334	−0.245	0.973	145.536	0.000001	0.80809	11.94%	232
Pain	−0.846	0.860	−0.356	0.630	44.100	0.00015	0.69772	1.44%	209
General self-image	0.466	0.157	−0.335	0.729	16.818	0.051647	0.71540	6.76%	204
Self-image postop	−1.695	0.611	−0.204	0.724	40.707	0.00000	0.30969	n/a	154
Function postop	−0.212	0.031	−0.267	0.346	24.428	0.000067	0.66911	n/a	146
General function	−1.846	1.169	−0.367	0.374	26.844	0.00006	0.03814	n/a	228
Function-activity	0.342	1.254	−0.189	0.639	11.958	0.007529	0.50185	n/a	110
Satisfaction with surgery	0.055	0.986	−0.256	0.646	26.907	0.001449	0.62845	0.625	162

Table 3 Rasch model fit statistics for the adjusted pre- and post-operative scores

Analysis	Item fit residual		Person fit residual		Chi square interaction		PSI	*t* tests (CI)	*n*
	Mean	SD	Mean	SD	Value	*p*			
Preop score	0.002	0.647	−0.225	0.861	26.092	0.097672	0.66616	2.25%	232
Postop score	−0.066	0.929	−0.292	0.865	17.913	0.267272	0.69007	3.7%	189

Table 4 Item fit residuals for the pre- and postop scores with rescored response options

Item	FitResid	ChiSq	Prob	Rescored to
Preoperative items				
2—pain over last month	−0.253	2.962	0.397472	01223
3—feelings toward back	−1.006	3.381	0.387640	
4—level of activity	−0.681	7.530	0.056806	00112
7—level of work activity	−0.553	6.235	0.100726	00112
14—feel attractive	0.147	2.607	0.456190	01123
15—self-image	0.770	4.581	0.205169	01234
Postoperative items				
16—changes in function	−1.159	2.620	0.453940	
17—enjoy sports/hobbies	0.054	3.948	0.267083	
19—confidence	1.329	3.328	0.343799	
20—others view	0.047	2.550	0.466409	00011
23—looks	−0.600	5.467	0.140635	00123

model fit (Table 2). The items for "general self-image" form a unidimensional set of questions on the other hand with 6.76% of *t* tests significant, which is above the 5% limit but within the lower bound of the confidence interval in a binomial distribution. For four domains, tests for unidimensionality could not be carried out, because RUMM2020 does not allow its analysis of domains of not more than three questions. In total, in three out of the seven domains unidimensionality could be shown (Table 2). None of these domains show fit to the Rasch model, however, and summating them into a total score is therefore not supported. Due to the multidimensional nature of the seven domains and because they show largely misfit to the Rasch model, profound changes have to be made to form a score.

Model fit and unidimensionality for pre- and postoperative summated scores

After rescoring all items with disordered thresholds, the preoperative items (items 1–15) were combined and analyzed as one group probing whether a summated preoperative score is possible. The full preop score is unidimensional with only 4.4% of the *t* tests significant, but shows misfit to the Rasch model ($p < 0.000001$). Further analysis revealed that items 2, 3, 4, 7, 14 and 15 formed a linear ($p = 0.097672$) and unidimensional (2.25% of significant *t* tests) subset (fit residuals and statistics shown in Table 3, preop score). The individual item fit residuals for this subset of questions are given in Table 4. There was no local dependence for the remaining items. The person separation index (PSI) as a measure of internal consistency reliability was 0.64. The fit residuals of the preoperative items which can be added into a summated score are given in Table 4. Responsiveness could be demonstrated for this subset of items. The score was calculated on a scale ranging from 0 to 18. Pre-operatively the total score was 10.56 ± 3.36 versus 13.78 ± 2.59 (58.6 ± 18.7 vs. $76.6 \pm 14.4\%$) postoperatively after 24 months. Analysis by ANOVA in RUMM2020 reveals a significant difference between the pre- and postoperative values ($p < 0.0001$) indicating the sensitivity to change of the subset of items.

Analysis of the postoperative items (items 16–24) combined revealed multidimensionality (6.28% of *t* tests significant) as well as model misfit ($p = 0.000027$). Items 19, 21, 22 and 24 introduced model misfit as was indicated by the fit residuals and were omitted. The final analysis showed model fit ($p = 0.267272$) and unidimensionality with 3.7% *t* tests significant (Table 3). No local dependence of items could be observed. The PSI was 0.69. Fit residuals for the postoperative items are listed in Table 4.

Analysis of differential item functioning (DIF)

All groups of patients should respond to the questions in the same way reflecting the underlying level of discomfort

relating to the pathology being probed. If respondents with the same level of discomfort are more likely to score higher or lower on an item, it shows differential item functioning (DIF). Only if it can be shown that DIF is not present, items can be added and the score used for comparison of the conditions DIF was tested for. In this investigation, person factors such as sex, age, whether the questionnaire was filled in pre- or post-operatively and whether it was a fusion from anterior, posterior or combined, were recorded. In each of these groups no DIF could be demonstrated. AIS patients with any of the above person factors respond in the same way to the questions asked in the shortened version of the SRS-24 questionnaire, putting them on the same linear scale and allowing direct comparison of their scores.

Discussion

The SRS-24 questionnaire has been introduced and validated using traditional psychometric methods, so called classic test-theory [3]. Reliability and external validity was shown and the questionnaire therefore found to be reproducible and to reflect the level of discomfort and satisfaction of scoliosis patients. While sensitivity to change or so-called responsiveness was not shown, the ultimate goal of the questionnaire is to calculate cross-sectional and longitudinal change scores. In order to calculate change scores, the requirements for internal construct validity such as unidimensionality and linearity have to be fulfilled [10].

In the present study, internal construct validity of the SRS-24 questionnaire was investigated. As fit to the Rasch model could not be demonstrated, a more detailed analysis was carried out and profound changes made to the scale. As a source of model misfit, i.e. non-linearity, disordered response options were found for several questions, which means that too many response options were presented in the questionnaire and that high levels of discomfort not necessarily endorse high responses for the specific question. This is particularly obvious in items 1 and 2 in which a redundant Likert-scale of 1–9 is reduced to 5–1 introducing disordering of response options. Only rescoring and collapsing the response options to four could establish a sequential order. The response options of all items of the SRS-24 questionnaire with disordered responses need to be changed to get ordered response options. Analysis of unidimensionality revealed a largely multidimensional scale which reflects how the scale was originally designed. Multidimensionality does not support the calculation of a total score. As most of the single dimensions are not unidimensional, their use as a subscore is not supported either. Misfitting questions therefore had to be removed and regrouped into pre- and postoperative items to obtain linearity and unidimensionality. The resulting scale comprises six preoperative and five postoperative items. Adding these two subsets again gives a multidimensional construct indicating that the pre- and postoperative items reflect two separate dimensions in scoliosis patients, which should not be combined into a total score. By themselves they are unidimensional and can therefore be used as standalone subscales. For clinical use of the pre- and postoperative subsets, rescoring of the questions as indicated in Table 4 has to be taken into account. Some response options in the Likert-type format have to be combined to offer fewer responses which ultimately results in ordered responses to each question.

While reliability and external validity were shown for the SRS-24 questionnaire, sensitivity to change or so-called responsiveness was not demonstrated initially [3]. Later studies indicated that especially the pain subscale may be useful for longitudinal assessment before and after surgery [5]. In this study, the preoperative items were reduced to a number of six items which proved to be sensitive to change from pre- to 24 months postoperative. Those six items consisting of questions regarding pain, activity and self-image appear to form a linear dimension in AIS patients which improves significantly postoperatively.

Conclusion

In summary, the SRS-24 questionnaire has been shown not to fulfill modern psychometric properties such as linearity and unidimensionality, which are required if a total score is to be calculated. The original questionnaire gives too many response options for several items and may therefore indicate wrong levels of discomfort for the individual patient. Profound changes to the questionnaire, such as reducing response options as well as removing and regrouping items into pre- and postoperative subscales had to be made to obtain linear and multidimensional scales. For clinicians we propose not to report the SRS-24 as a total score and separate pre- and postoperative items using the subsets identified in this study. Both subscores using the pre- and postoperative items can be derived from existing SRS-24 scores, while the six preoperative items have been shown to be sensitive to change during follow-up.

Conflict of interest The authors declare no competing interest.

References

1. Bland JM, Altman DG (1995) Multiple significance tests: the bonferroni method. BMJ 310(6973):170
2. Comer CM, Conaghan PG, Tennant A (2011) Internal construct validity of the swiss spinal stenosis questionnaire: Rasch analysis of a disease-specific outcome measure for lumbar spinal stenosis. Spine (Phila Pa 1976) 36(23):1969–1976

3. Haher TR, Gorup JM, Shin TM, Homel P, Merola AA, Grogan DP, Pugh L, Lowe TG, Murray M (1999) Results of the scoliosis research society instrument for evaluation of surgical outcome in adolescent idiopathic scoliosis: a multicenter study of 244 patients. Spine (Phila Pa 1976) 24(14):1435–1440
4. Luce R, Tukey J (1964) Simultaneous conjoint measurement: a newtype of fundamental measurement. J Mathemat Psychol 1:1–27
5. Merola AA, Haher TR, Brkaric M, Panagopoulos G, Mathur S, Kohani O, Lowe TG, Lenke LG, Wenger DR, Newton PO, Clements DH 3rd, Betz RR (2002) A multicenter study of the outcomes of the surgical treatment of adolescent idiopathic scoliosis using the scoliosis research society (srs) outcome instrument. Spine (Phila Pa 1976) 27(18):2046–2051
6. Pallant JF, Keenan AM, Misajon R, Conaghan PG, Tennant A (2009) Measuring the impact and distress of osteoarthritis from the patients' perspective. Health Qual Life Outcomes 7:37. doi: 10.1186/1477-7525-7-37
7. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago
8. Rothenfluh DA, Reedwisch D, Muller U, Ganz R, Tennant A, Leunig M (2008) Construct validity of a 12-item womac for assessment of femoro-acetabular impingement and osteoarthritis of the hip. Osteoarthr Cartil 16(9):1032–1038
9. Streiner D, Norman G (1989) Health measurement scales. Oxford University Press, Oxford
10. Svensson E (2001) Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 33(1):47–48
11. Wright B (1997) A history of social science and measurement. Educ Meas Issues Pract Winter 16:33–45